

J.P.T.O. / J.P.T.O. CO. J.

日 本 国 特 許 庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日 2 0 0 2 年 1 1 月 2 8 日
Date of Application:

出 願 番 号 特 願 2 0 0 2 - 3 4 5 9 8 8
Application Number:

[ST. 10/C]: [J P 2 0 0 2 - 3 4 5 9 8 8]

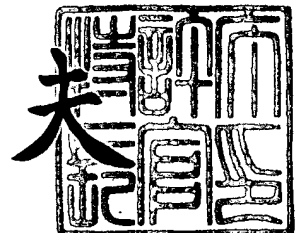
出 願 人 沖 電 気 工 業 株 式 会 社
Applicant(s):

CERTIFIED COPY OF
PRIORITY DOCUMENT

2 0 0 3 年 1 1 月 2 8 日

特許庁長官
Commissioner,
Japan Patent Office

今 井 康 夫



【書類名】 特許願

【整理番号】 KT000468

【提出日】 平成14年11月28日

【あて先】 特許庁長官 太田 信一郎 殿

【国際特許分類】 G06F 17/28

【発明者】

 【住所又は居所】 東京都港区虎ノ門 1 丁目 7 番 1 2 号 沖電気工業株式会
社内

 【氏名】 介弘 達哉

【特許出願人】

 【識別番号】 000000295

 【氏名又は名称】 沖電気工業株式会社

【代理人】

 【識別番号】 100095957

 【弁理士】

 【氏名又は名称】 亀谷 美明

 【電話番号】 03-5919-3808

【選任した代理人】

 【識別番号】 100096389

 【弁理士】

 【氏名又は名称】 金本 哲男

 【電話番号】 03-3226-6631

【選任した代理人】

 【識別番号】 100101557

 【弁理士】

 【氏名又は名称】 萩原 康司

 【電話番号】 03-3226-6631

【手数料の表示】

【予納台帳番号】 040224

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【包括委任状番号】 9707549

【包括委任状番号】 9707550

【包括委任状番号】 9707551

【ブルーフの要否】 要

【書類名】 明細書

【発明の名称】 複数言語文書の対応付けシステム、複数言語文書の対応付け方法、及びプログラム並びにプログラムを記録した記録媒体

【特許請求の範囲】

【請求項 1】 n 種 (n は 2 以上の自然数) の言語の文書を対応付けるシステムであって、

各言語の文書を単語毎に分割する形態素解析手段と、

前記 n 種の言語の文書のうちの 2 種を選択する手段と、

前記選択された 2 種の言語の文書の評価関数を計算する手段と、

前記 2 種の言語の文書の評価結果によって、前記 n 種の言語の文書を対応付ける手段と、

を含むことを特徴とする、複数言語文書の対応付けシステム。

【請求項 2】 前記形態素解析手段が、

各言語の文書を文毎に分割する手段と、分割された各文をさらに単語毎に分割する手段とからなることを特徴とする、請求項 1 に記載の複数言語文書の対応付けシステム。

【請求項 3】 前記 n 種の言語の文書のうちの 2 種を選択する手段が、

前記 n 種の言語の文書を任意の順序で並べたときに、 k 番目と $k+1$ 番目 (k は、1 から $n-1$ までの自然数) の、 $n-1$ 通りの組合せを選択することを特徴とする、請求項 1 または 2 のうちのいずれか 1 項に記載の複数言語文書の対応付けシステム。

【請求項 4】 前記 n 種の言語の文書のうちの 2 種を選択する手段が、

$n(n-1)/2$ 通りの全ての組合せを選択することを特徴とする、請求項 1 または 2 のうちのいずれか 1 項に記載の複数言語文書の対応付けシステム。

【請求項 5】 さらに、前記評価関数で計算した結果を保持する計算結果管理手段を含むことを特徴とする、請求項 1, 2, 3, または 4 のうちのいずれか 1 項に記載の複数言語文書の対応付けシステム。

【請求項 6】 前記評価関数が、次式；

$$h(x, y) = 2 \times f_m(x, y) / (f_j(x) + f_j(y))$$

で表されることを特徴とする、請求項 1, 2, 3, 4, または 5 のうちのいずれか 1 項に記載の複数言語文書の対応付けシステム。

但し、 $h(x, y)$ は、評価関数、

x は、一方の言語の文、

y は、他方の言語の文、

$f_m(x, y)$ は、文 x と文 y の中で対応の付いた自立語の数、

$f_j(x)$ は、文 x 中の自立語の数、

$f_j(y)$ は、文 y 中の自立語の数、

である。

【請求項 7】 さらに、 n 種のうちのいずれかの 3 種以上の言語の文書の対応付けに不整合が生じたときに、前記不整合の箇所を表示する手段を含むことを特徴とする、請求項 1, 2, 3, 4, 5, または 6 のうちのいずれか 1 項に記載の複数言語文書の対応付けシステム。

【請求項 8】 前記評価関数を計算する手段は、前記評価関数の和が最大になるように最適化しながら対応付けを行うことを特徴とする、請求項 1, 2, 3, 4, 5, 6, または 7 のうちのいずれか 1 項に記載の複数言語文書の対応付けシステム。

【請求項 9】 さらに、言語間の類似度データを調べながら、対応付けの正解率の高い言語対を指示する手段を含むことを特徴とする、請求項 1, 2, 3, 4, 5, 6, 7, または 8 のうちのいずれか 1 項に記載の複数言語文書の対応付けシステム。

【請求項 10】 n 種 (n は 2 以上の自然数) の言語の文書を対応付ける方法であって、

各言語の文書を単語毎に分割する形態素解析段階と、

前記 n 種の言語の文書のうちの 2 種を選択する段階と、

前記選択された 2 種の言語文書の評価関数を計算する段階と、

前記 2 種の言語の文書の評価結果によって、 n 種の言語の文書を対応付ける段階と、

を含むことを特徴とする、複数言語文書の対応付け方法。

【請求項 11】 前記形態素解析段階が、

各言語の文書を文毎に分割する段階と、分割された各文をさらに単語毎に分割する段階とからなることを特徴とする、請求項 10 に記載の複数言語文書の対応付け方法。

【請求項 12】 前記 n 種の言語の文書のうちの 2 種を選択する段階が、

前記 n 種の言語の文書を任意の順序で並べたときに、 k 番目と $k+1$ 番目 (k は、1 から $n-1$ までの自然数) の、 $n-1$ 通りの組合せを選択することを特徴とする、請求項 10 または 11 のうちのいずれか 1 項に記載の複数言語文書の対応付け方法。

【請求項 13】 前記 n 種の言語の文書のうちの 2 種を選択する段階が、

$n(n-1)/2$ 通りの全ての組合せを選択することを特徴とする、請求項 10 または 11 のうちのいずれか 1 項に記載の複数言語文書の対応付け方法。

【請求項 14】 さらに、前記評価関数で計算した結果を保持する計算結果管理段階を含むことを特徴とする、請求項 10, 11, 12, または 13 のうちのいずれか 1 項に記載の複数言語文書の対応付け方法。

【請求項 15】 前記評価関数が、次式；

$$h(x, y) = 2 \times f_m(x, y) / (f_j(x) + f_j(y))$$

で表されることを特徴とする、請求項 10, 11, 12, 13, または 14 のうちのいずれか 1 項に記載の複数言語文書の対応付け方法。

但し、 $h(x, y)$ は、評価関数、

x は、一方の言語の文、

y は、他方の言語の文、

$f_m(x, y)$ は、文 x と文 y の中で対応のついた自立語の数、

$f_j(x)$ は、文 x 中の自立語の数、

$f_j(y)$ は、文 y 中の自立語の数、

である。

【請求項 16】 さらに、 n 種のうちのいずれかの 3 種以上の言語の文書の対応付けに不整合が生じたときに、前記不整合の箇所を表示する段階を含むことを特徴とする、請求項 10, 11, 12, 13, 14, または 15 のうちのいづ



れか 1 項に記載の複数言語文書の対応付け方法。

【請求項 17】 前記評価関数を計算する段階は、前記評価関数の和が最大になるように最適化しながら対応付けを行うことを特徴とする、請求項 10、11、12、13、14、15、または 16 のうちのいずれか 1 項に記載の複数言語文書の対応付け方法。

【請求項 18】 さらに、言語間の類似度データを調べながら、対応付けの正解率の高い言語対を指示する段階を含むことを特徴とする、請求項 10、11、12、13、14、15、16、または 17 のうちのいずれか 1 項に記載の複数言語文書の対応付け方法。

【請求項 19】 コンピュータに、請求項 10、11、12、13、14、15、16、17、または 18 のうちのいずれか 1 項に記載の複数言語文書の対応付け方法を行わせるステップを記述したことを特徴とする、プログラム。

【請求項 20】 コンピュータに、請求項 10、11、12、13、14、15、16、17、または 18 のうちのいずれか 1 項に記載の複数言語文書の対応付け方法を行わせるプログラムが記録されていることを特徴とする、記録媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、複数の言語で構成される文書間の文書対応付けシステムにかかり、特に、2 言語以上で記述された対訳文書の、文の対応付けを行う複数言語文書の対応付けシステム、複数言語文書の対応付け方法、この方法を行わせるプログラム、及びこのプログラムを記録した記録媒体に関する。

【0002】

【従来の技術】

海外に輸出される製品のマニュアルなどのように、複数の言語で同じ内容の文書を記述する場合が増えている。このような複数の言語文書の対訳の正確性を評価、担保等するため、これらの文の対応付けを行う需要も増えている。非特許文献 1 は、対訳文書の文の対応付けを、対訳辞書を利用したダイナミックプログラ

ミングで行う方法が記載されている。

【0003】

非特許文献1によれば、対応付けを行うには、文書を1文毎に区切り、さらにその文の形態素解析を行って、単語毎に分割する。そして、これらの単語の中から自立語を取り出し、対訳辞書を用いてそれぞれの文の中の自立語がどの程度対応しているか（どの程度意味内容が一致しているか）によって対応付けを評価する。評価では、例えば以下のような式を用いる。

【0004】

【数1】

$$h(x, y) = 2 \times f_m(x, y) / (f_j(x) + f_j(y))$$

【0005】

ここで

$h(x, y)$ は、評価関数,

x は、原文中の文（複数文の場合もある）,

y は、訳文中の文（複数文の場合もある）,

$f_m(x, y)$ は、文 x と文 y の中で対応の付いた自立語の数,

$f_j(x)$ は、文 x 中の自立語の数,

$f_j(y)$ は、文 y 中の自立語の数,

である。

【0006】

このような式による評価を行えば、文書の対応の割合が大きいほど評価関数 $h(x, y)$ の値は大きくなり（最大；1）、逆は小さくなる（最小；0）。この評価関数を文の先頭から調べていき、評価関数の和が最も大きくなる組合せを、対応付け問題の解とする。

【0007】

【非特許文献1】

宇津呂 武仁, 松本 裕治 共著「対訳辞書及び統計情報を用いた二言語対訳テキスト照合」（「コンピュータソフトウェア」岩波書店 vol.12 No. 5 Sep.1995 p.12(414)-p.21(423)）

【0008】**【発明が解決しようとする課題】**

しかしながら、上記方法では、通常の 2 言語の対訳文書の文の対応付けを、3 言語以上の文書の文の対応付けに適用する場合に、

- ・複数の辞書を使用するため、システムにかなりの量の記録領域を必要とする。
- ・評価の処理に時間がかかる。
- ・全ての言語間で、各言語対の対応の整合性をとるのが困難である。

などの問題がある。

【0009】

また、2 言語の対訳文書の対応付けに関しても、高精度での対応を自動的に付けるのは難しく、対応付けの結果を見ながらの人の手によるチェックや修正が必要であり、その作業時間が問題となっている。

【0010】

本発明は、従来の複数言語文書の対応付けシステムが有する上記問題点に鑑みてなされたものである。そして、本発明の目的は、英語－日本語－ドイツ語など、複数の言語でそれぞれ構成される文書間の文の対応付けを効率良く行うための、新規かつ改良された複数言語文書の対応付けシステム、及び複数言語文書の対応付け方法を提供することにある。

【0011】**【課題を解決するための手段】**

上記課題を解決するための本発明の複数言語文書の対応付けシステムは、 n 種（ n は 2 以上）の言語の文書を対応付けるシステムである。そして、各言語の文書を単語毎に分割する形態素解析手段と、 n 種の言語の文書のうちの 2 種を選択する手段と、選択された 2 種の言語文書の評価関数を計算する手段と、評価結果に応じて n 種の言語の文書を対応付ける手段とから構成される。

【0012】

ここで、各言語の文書を単語毎に分割する形態素解析手段は、各言語の文書を文毎に分割する手段と、分割された各文をさらに単語毎に分割する手段とからなってもよい。

【0013】

【発明の実施の形態】

以下に添付図面を参照しながら、本発明にかかる複数言語文書の対応付けシステム、複数言語文書の対応付け方法の好適な実施の形態について詳細に説明する。

【0014】

(第1の実施の形態)

図1は、第1の実施の形態にかかる複数言語文書の対応付けシステム100の構成を示す説明図である。複数言語文書の対応付けシステム100は、図1に示したように、文分割手段105と、形態素解析手段106と、評価関数計算手段107と、計算結果管理手段108と、対訳辞書データベース109により構成されている。この例では、各言語のファイル101～104が入力されて、対応タグ付きファイル110～113が出力される。以下、各構成要素につき詳細に説明する。

【0015】

英語ファイル101は、英語で記述された文書ファイル、日本語ファイル102は、日本語で記述された文書ファイル、ドイツ語ファイル103は、ドイツ語で記述された文書ファイル、中国語ファイル104は、中国語で記述された文書ファイルである。上記4つのファイルはそれぞれ同じ内容のことを述べており、それぞれが対訳形式になっている。

【0016】

文分割手段105は、文書ファイルを1文毎に分割する。例えば、英文であればピリオド「.」, 日本文なら句点「。」などで分割する。形態素解析手段106は、形態素解析処理を行い、文を単語毎に分割する。文分割手段105及び形態素解析手段106は、既存のものを適用できる。

【0017】

評価関数計算手段107は、最適な対応付けを見つけるために、与えられた評価関数を計算する。例えば、評価関数は、次式；

$$h(x, y) = 2 \times f_m(x, y) / (f_j(x) + f_j(y))$$

で表される。ここで、 $h(x, y)$ は、評価関数であり、 x は、一方の言語の文（原文）であり、 y は、他方の言語の文（訳文）であり、 $f_m(x, y)$ は、文 x と文 y の中で対応の付いた自立語の数であり、 $f_j(x)$ は、文 x 中の自立語の数であり、 $f_j(y)$ は、文 y 中の自立語の数である。

【0018】

計算結果管理手段 108 は、評価関数計算手段が計算した結果を保持し、既出の評価関数計算が再び到来したときに保持した結果を出力し、同じ計算を何度も行わないようにする。

【0019】

対訳辞書データベース 109 は、対応付けをするための原文の単語を引くと訳文の語が 1 つまたは複数あるような辞書である。例えば、原文が英語、訳文が日本語の場合、英和辞典に相当する。

【0020】

対応タグ付英語ファイル 110 は、英語ファイルに他の文書のどの文に対応しているかを示すタグを付与したものである。対応タグ付き日本語ファイル 111、対応タグ付きドイツ語ファイル 112、及び対応タグ付中国語ファイル 113 も同様に、元のファイルに文の対応を示すためのタグを付与したものである。

【0021】

本実施の形態にかかる複数言語文書の対応付けシステム 100 は、以上のように構成されている。次に、図 2 を参照しながら、複数言語文書の対応付けシステム 100 の動作を説明する。

【0022】

図 2 は、第 1 の実施の形態の複数言語文書の対応付けシステム 100 の動作を示すフローチャートである。ステップ S10 では文分割手段によって一方の文（原文）ファイルと他方の文（訳文）ファイルの文分割を行う。そして、対応付けをどこまで行ったかを示すカウンタ N を、0 にセットする。

ステップ S11 では、カウンタ N をインクリメント（+1）する。

ステップ S12 では、対応付けを行う言語の数がカウンタ N と等しいかどうかを比較する。もし等しければ、ステップ S17 に行く。

【 0 0 2 3 】

ステップ S 1 3 では、対応付けを行う言語を N 番目と N + 1 番目にセットする。

ステップ S 1 4 では、評価関数計算手段がセットされた言語に対して文の対応付けを行う。

ステップ S 1 5 では、対応付けを行った結果に対して、対応する文同士に双方向リンクを張る。

【 0 0 2 4 】

ステップ S 1 6 では、2 対 1，3 対 1 などの複数文の対応になった文に対してマーク付けを行う。これらのマーク付けされた文の組は、次に対応付けを行う場合はそれを主文とみなして処理する。

ステップ S 1 7 では、対応付けを行っていない言語同士の文に対して、他の言語同士の対応付け結果を利用して、リンクを張る。

【 0 0 2 5 】

以上の処理を、図 1 の 4 言語の対応付けを行う場合に関して説明する。この例では、英語が 1 番目、日本語が 2 番目、ドイツ語が 3 番目、中国語が 4 番目の言語に相当する。

【 0 0 2 6 】

まず、4 つの言語 ($n = 4$) それぞれを文分割手段によって一文毎に分割する。次に、文の対応付けを行う。英語と日本語の対応付けを英日対訳辞書を使って、日本語とドイツ語の対応付けを日独対訳辞書を使って、ドイツ語と中国語の対応付けを独中対訳辞書を使ってそれぞれ行う。これにより、日本語－英語間、日本語－ドイツ語間、ドイツ語－中国語間の ($n - 1$) 通りの文同士のリンクが生成される。

【 0 0 2 7 】

さらに、対応のついていない言語同士（ここでは、日本語－中国語、英語－ドイツ語、英語－中国語）の文のリンクを張ることによって、すべての言語間の文の対応をとることができる。

【 0 0 2 8 】

以上説明したように、本実施の形態によれば、対応付けの精度は多少落ちるが、少ない記録容量で時間もあまりかからずに効率良く文の対応をとることができる。

【0 0 2 9】

(第 2 の実施の形態)

図 3 に、第 2 の実施の形態の複数言語文書の対応付けシステムの構成を示す。英語ファイル 2 0 1 は、英語で記述された文書ファイル、日本語ファイル 2 0 2 は、日本語で記述された文書ファイル、ドイツ語ファイル 2 0 3 は、ドイツ語で記述された文書ファイル、中国語ファイル 2 0 4 は、中国語で記述された文書ファイルである。上記 4 つのファイルは、それぞれ同じ内容が記述されており、それぞれが対訳形式になっている。

【0 0 3 0】

文分割手段 2 0 5 は、文書ファイルを 1 文毎に分割する。英文であればピリオド「.」，日本文なら句点「。」などで分割する。形態素解析手段 2 0 6 は、形態素解析処理を行い、文を単語毎に分割する。文分割手段 2 0 5 及び形態素解析手段 2 0 6 は、既存のものを適用できる。評価関数計算手段 2 0 7 は、最適な対応付けを見つけるために、与えられた評価関数を計算する。評価関数は、例えば第 1 の実施の形態で示したものが適用できる。

【0 0 3 1】

計算結果管理手段 2 0 8 は、評価関数計算手段が計算した結果を保持し、既出の評価関数計算が再び到来したときに保持した結果を出力し、同じ計算を何度も行わないようにする。対訳辞書データベース 2 0 9 は、対応付けをするための辞書で、原文の単語を引くと訳文の語が 1 つまたは複数あるような辞書である。原文が英語、訳文が日本語の場合、英和辞典に相当する。

【0 0 3 2】

対応タグ付英語ファイル 2 1 0 は、英語ファイルに他の文書のどの文に対応しているかを示すタグを付与したものである。対応タグ付き日本語ファイル 2 1 1，対応タグ付きドイツ語ファイル 2 1 2，及び対応タグ付中国語ファイル 2 1 3 も同様に、元のファイルに文の対応を示すためのタグを付与したものである。

【0033】

相違箇所表示手段 220 は、対応付け結果に不整合があった場合に、その不整合箇所を表示し、ユーザに修正させる機能をもつ。不整合とは、例えば、英語の文 E_n と日本語の J_n が対応していて、日本語の文 J_n とドイツ語の文 D_n が対応しているときに、英語とドイツ語の対応結果をみると、英文 E_n とドイツ文 D_n とが対応していないような場合である。

【0034】

図 4 は、本実施の形態の複数言語文書の対応付けシステム 200 の動作を示すフローチャートである。

ステップ S20 では、文分割手段によって一方の文（原文）ファイルと他方の文（訳文）ファイルの文分割を行う。そして、対応付けをどこまで行ったかを示すカウンタ N と M を、1 にセットする。

ステップ S21 では、対応付けを行う言語の数がカウンタ N と等しいかどうかを比較する。もし等しければ、ステップ S27 に行く。

ステップ S22 では、カウンタ M をインクリメントし、N の値を M+1 にする。

ステップ S23 では、対応付けを行う言語の数がカウンタ M と等しいかどうかを比較する。もし等しければ、S28 に行く。

【0035】

ステップ S24 では、対応付けを行う言語を M 番目と N 番目にセットする。

ステップ S25 では、評価関数計算手段がセットされた言語に対して文の対応付けを行う。

ステップ S26 では、対応付けを行った結果に対して、対応する文同士に双方向リンクを張る。

ステップ S27 では、N をインクリメントする。

ステップ S28 では、文の対応に不整合がある部分を表示しユーザに修正させる。

ステップ S29 では、ユーザの修正に応じて、対応付けのリンクを張り直す。

このようにして、n 種の言語の文に対して、全ての組合せ（この例では、言語

の種類 $n = 4$ で、 $n(n-1)/2 = 6$ 通り) の対応付けを行う。

【0036】

以上説明したように、本実施の形態によれば、ユーザが修正することが必須であるが、高精度の対応付けが効率良く実現できる。

【0037】

(第3の実施の形態)

図5に、第3の実施の形態の複数言語文書の対応付けシステムの構成を示す。英語ファイル301は、英語で記述された文書ファイル、日本語ファイル302は、日本語で記述された文書ファイル、ドイツ語ファイル303は、ドイツ語で記述された文書ファイル、中国語ファイル304は、中国語で記述された文書ファイルである。上記4つのファイルは、それぞれ同じ内容が記述されており、それぞれが対訳形式になっている。

【0038】

文分割手段305は、文書ファイルを1文毎に分割する。英文であればピリオド「.」, 日本文なら句点「。」などで分割する。形態素解析手段306は、形態素解析処理を行い、文を単語毎に分割する。文分割手段305及び形態素解析手段306は、既存のものを適用できる。評価関数計算手段307は、最適な対応付けを見つけるために、与えられた評価関数を計算する。評価関数は、例えば第1の実施の形態で示したものが適用できる。

【0039】

計算結果管理手段308は、評価関数計算手段が計算した結果を保持し、既出の評価関数計算が再び到来したときに保持した結果を出力し、同じ計算を何度も行わないようにする。

【0040】

対訳辞書データベース309は、対応付けをするための辞書で、原文の単語を引くと訳文の語が1つまたは複数あるような辞書である。原文が英語、訳文が日本語の場合英和辞典に相当する。

【0041】

対応タグ付英語ファイル310は、英語ファイルに他の文書のどの文に対応し

ているかを示すタグを付与したものである。対応タグ付き日本語ファイル 311, 対応タグ付きドイツ語ファイル 312, 及び対応タグ付中国語ファイル 313 も同様に, 元のファイルに文の対応を示すためのタグを付与したものである。

【0042】

図6は, 本実施の形態の複数言語文書の対応付けシステム 300 の動作を示すフローチャートである。

ステップ S30 では, 文分割手段によって一方の文 (原文) ファイルと他方の文 (訳文) ファイルの文分割を行う。そして, 対応付けをどこまで行ったかを示すカウンタ N と M を, 1 にセットする。

ステップ S31 では, 対応付けを行う言語の数がカウンタ N と等しいかどうかを比較する。もし等しければ, ステップ S37 に行く。

【0043】

ステップ S32 では, カウンタ M をインクリメントし, N の値を M+1 にする。

ステップ S33 では, 対応付けを行う言語の数がカウンタ M と等しいかどうかを比較する。もし等しければ, ステップ S37 に行く。

【0044】

ステップ S34 では, 対応付けを行う言語を M 番目と N 番目にセットする。

ステップ S35 では, 評価関数計算手段がセットされた言語に対して評価関数を計算する。

ステップ S36 では, N をインクリメントする。

ステップ S37 では, 対応付けのポイントの和が最も大きくなるような文の組を選択する。

ステップ S38 では, 対応する文同士に双方向リンクを張る。

【0045】

以上の処理を, 図5の4言語 ($n=4$) の対応付けを行う場合に関して説明する。例では, 英語が1番目, 日本語が2番目, ドイツ語が3番目, 中国語が4番目の言語に相当する。

【0046】

まず、4つの言語それぞれを文分割手段によって一文毎に分割する。次に、すべての文書の組の評価関数を計算する。この場合、英語－日本語、英語－ドイツ語、英語－中国語、日本語－ドイツ語、日本語－中国語、ドイツ語－中国語の6つの評価関数を計算する。

【0 0 4 7】

次に、対応付けポイントの和が最も大きくなるように対応をとっていく。この対応は4言語まとめて同時に行われる。例えば、英文1文、日本文1文、ドイツ文2文、中国文1文の評価ポイントは、英文と日本文の1文対1文、英文とドイツ文の1文対2文、英文と中国文の1文対1文、日本文とドイツ文の1文対2文、日本文と中国文の1文対1文、ドイツ文と中国文の2文対1文、の評価ポイントの和となる。この計算を続け、評価ポイントの和の和が最も大きくなったものを対応付けの正解とする。

【0 0 4 8】

以上説明したように、本実施の形態によれば、時間はかなりかかるが高精度の対応付けが効率良く実現できる。

【0 0 4 9】

(第4の実施の形態)

図7に、第4の実施の形態の複数言語文書の対応付けシステムの構成を示す。英語ファイル401は英語で記述された文書ファイル、日本語ファイル402は日本語で記述された文書ファイル、ドイツ語ファイル403はドイツ語で記述された文書ファイル、中国語ファイル404は中国語で記述された文書ファイルである。上記4つのファイルはそれぞれ同じ内容が記述されており、それぞれが対訳形式になっている。

【0 0 5 0】

文分割手段405は、文書ファイルを1文毎に分割する。英文であればピリオド「.」，日本文なら句点「。」などで分割する。形態素解析手段406は、形態素解析処理を行い、文を単語毎に分割する。文分割手段405及び形態素解析手段406は、既存のものを適用できる。評価関数計算手段407は、最適な対応付けを見つけるために、与えられた評価関数を計算する。評価関数は、例えば

第 1 の実施の形態で示したものが適用できる。

【0 0 5 1】

計算結果管理手段 4 0 8 は、評価関数計算手段が計算した結果を保持し、既出の評価関数計算が再び到来したときに保持した結果を出力し、同じ計算を何度も行わないようにする。対訳辞書データベース 4 0 9 は、対応付けをするための辞書で、原文の単語を引くと訳文の語が 1 つまたは複数あるような辞書である。原文が英語、訳文が日本語の場合英和辞典に相当する。

【0 0 5 2】

対応タグ付英語ファイル 4 1 0 は、英語ファイルに他の文書のどの文に対応しているかを示すタグを付与したものである。対応タグ付き日本語ファイル 4 1 1、対応タグ付きドイツ語ファイル 4 1 2、及び対応タグ付中国語ファイル 4 1 3 も同様に、元のファイルに文の対応を示すためのタグを付与したものである。

【0 0 5 3】

言語類似度データ 4 2 0 は、言語同士の文法などがどれだけ似ているかを数値化したものである。類似度が高いほど文の対応付けの程度も向上する。それぞれの言語対の類似度の値が、例えば表形式などで記録されている。

【0 0 5 4】

図 8 は、本実施の形態の複数言語文書の対応付けシステム 4 0 0 の動作を示すフローチャートである。

ステップ S 4 0 では、文分割手段によって一方の文ファイルと他方の文ファイルの文分割を行う。

対応付けをどこまで行ったかを示すカウンタ N を、0 にセットする。

ステップ S 4 1 では、カウンタ N をインクリメントする。

ステップ S 4 2 では、対応付けを行う言語の数がカウンタ N と等しいかどうかを比較する。もし等しければ、終了する。

【0 0 5 5】

ステップ S 4 3 では、言語類似度が最も高く、まだ選択されていない言語対を選択し、選択済みのマークをつけておく。

ステップ S 4 4 では、言語対に文対応のリンクが張られているかどうかを調べ

る。リンクがすでに張られていれば、ステップ S 4 3 に行く。

【0056】

ステップ S 4 5 では、評価関数計算手段が選択された言語に対して文の対応付けを行う。

ステップ S 4 6 では、対応付けを行った結果に対して、対応する文同士に双方向リンクを張る。

ステップ S 4 7 では、2 対 1、3 対 1 などの複数文の対応になった文に対してマーク付けを行う。これらのマーク付けされた文の組は次に対応付けを行う場合はそれを 1 文とみなして処理する。

ステップ S 4 8 では、間接的に対応のついた言語に対してリンクを張る。例えば、英語－日本語、英語－ドイツ語の対応がとれたとすると、日本語－ドイツ語間にも文対応のリンクを張る。

【0057】

以上説明したように、本実施の形態によれば、言語類似度データを用意する必要があるが、高速に精度の高い対応付けが効率良く実現できる。

【0058】

上記の 4 つの実施の形態の速度、精度、使用する記録容量を比較すると、表 1 のようになる。表 1 において、「◎」は優良、「○」は良好、「△」は普通である。

【0059】

【表 1】

実施の形態	速度	精度	記録容量	その他
1	◎	△	◎	
2	○	◎	△	ユーザの修正が必須
3	△	◎	△	
4	◎	○	○	言語類似度データが必要

【0060】

以上、添付図面を参照しながら本発明にかかる複数言語文書の対応付けシステ

ム、及び複数言語文書の対応付け方法の好適な実施形態について説明したが、本発明はかかる例に限定されない。当業者であれば、特許請求の範囲に記載された技術的思想の範疇内において各種の変更例または修正例に想到し得ることは明らかであり、それらについても当然に本発明の技術的範囲に属するものと了解される。

【0061】

例えば、上記第1～第4実施の形態では、英語、日本語、ドイツ語、中国語の対応付けを示したが、対訳辞書を変えることによって、どんな言語同士の対応もとることができる。また、4言語（ $n=4$ ）の例を示したが、2言語以上であれば何言語の対応付けにも対応できる。第2、第3の実施の形態では言語数が増えてくると処理時間が非常に遅くなるおそれがあるが、計算する対応組の数を減らすことによって対応できる。

【0062】

なお、本発明の複数言語文書の対応付け方法は、プログラムに記述することもでき、本発明の複数言語文書の対応付け方法を記述したプログラムは、記録媒体に記録することができる。

【0063】

【発明の効果】

以上説明したように、本発明によれば、複数の言語で構成される文書間の文の対応付けを効率良く行う複数言語文書の対応付けシステムが提供できた。

【図面の簡単な説明】

【図1】

第1の実施の形態にかかる複数言語文書の対応付けシステムの構成を示す説明図である。

【図2】

図1の複数言語文書の対応付けシステムの動作を示すフローチャートである。

【図3】

第2の実施の形態にかかる複数言語文書の対応付けシステムの構成を示す説明図である。

【図 4】

図 3 の複数言語文書の対応付けシステムの動作を示すフローチャートである。

【図 5】

第 3 の実施の形態にかかる複数言語文書の対応付けシステムの構成を示す説明図である。

【図 6】

図 5 の複数言語文書の対応付けシステムの動作を示すフローチャートである。

【図 7】

第 4 の実施の形態にかかる複数言語文書の対応付けシステムの構成を示す説明図である。

【図 8】

図 7 の複数言語文書の対応付けシステムの動作を示すフローチャートである。

【符号の説明】

100, 200, 300, 400	複数言語文書の対応付けシステム
101, 201, 301, 401	英語ファイル
102, 202, 302, 402	日本語ファイル
103, 203, 303, 403	ドイツ語ファイル
104, 204, 304, 404	中国語ファイル
105, 205, 305, 405	文分割手段
106, 206, 306, 406	形態素解析手段
107, 207, 307, 407	評価関数計算手段
108, 208, 308, 407	計算結果管理手段
109, 209, 309, 409	対訳辞書データベース
110, 210, 310, 410	対応タグ付英語ファイル
111, 211, 311, 411	対応タグ付日本語ファイル
112, 212, 312, 412	対応タグ付独語ファイル
113, 213, 313, 413	対応タグ付中国語ファイル
114, 214, 314, 414	英日対訳辞書
115, 215, 315, 415	日独対訳辞書

1 1 6, 2 1 6, 3 1 6, 4 1 6 独中対訳辞書

2 1 7, 3 1 7, 4 1 7 英独対訳辞書

2 1 8, 3 1 8, 4 1 8 英中対訳辞書

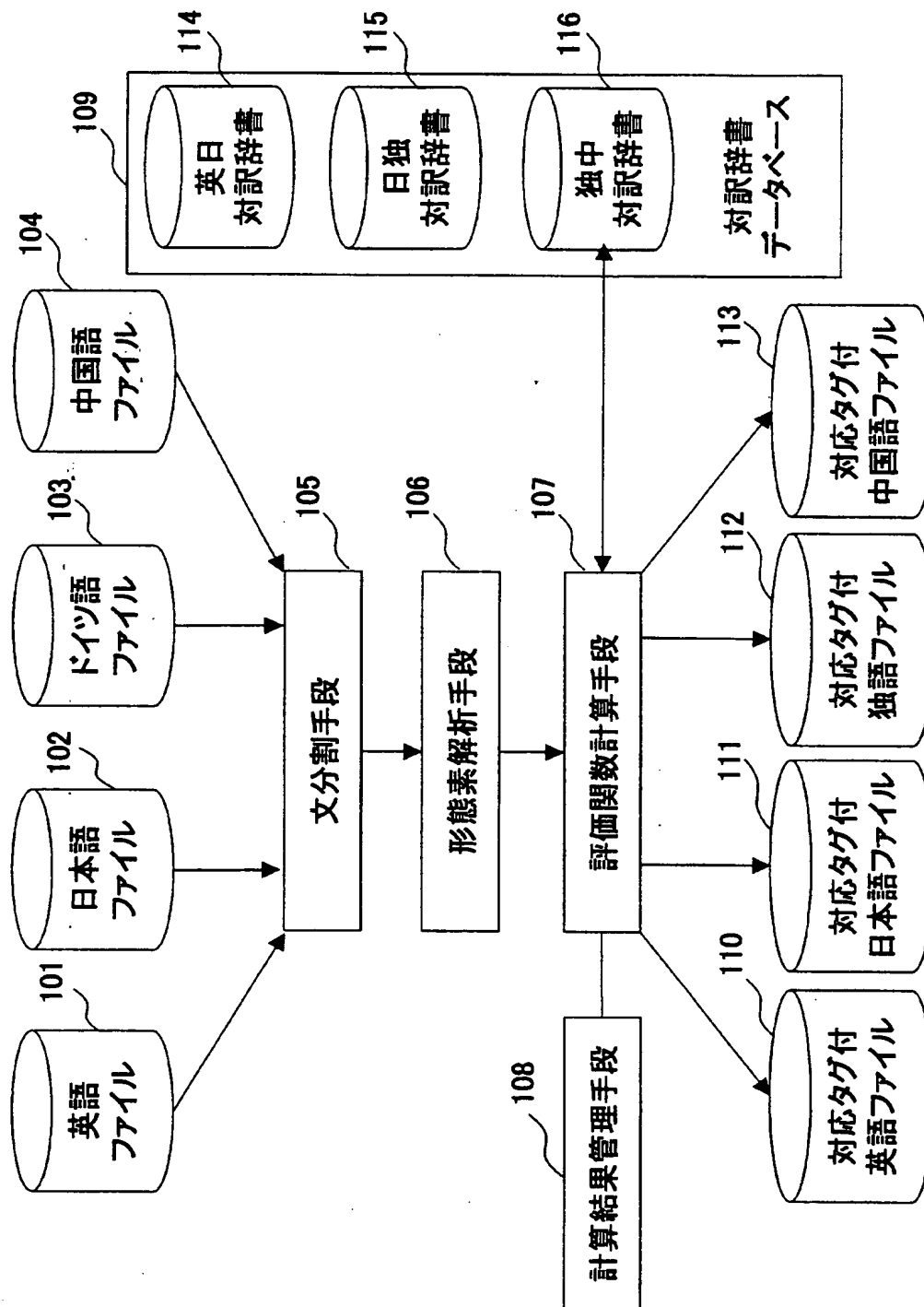
2 1 9, 3 1 9, 4 1 9 日中対訳辞書

2 2 0 相違箇所表示手段

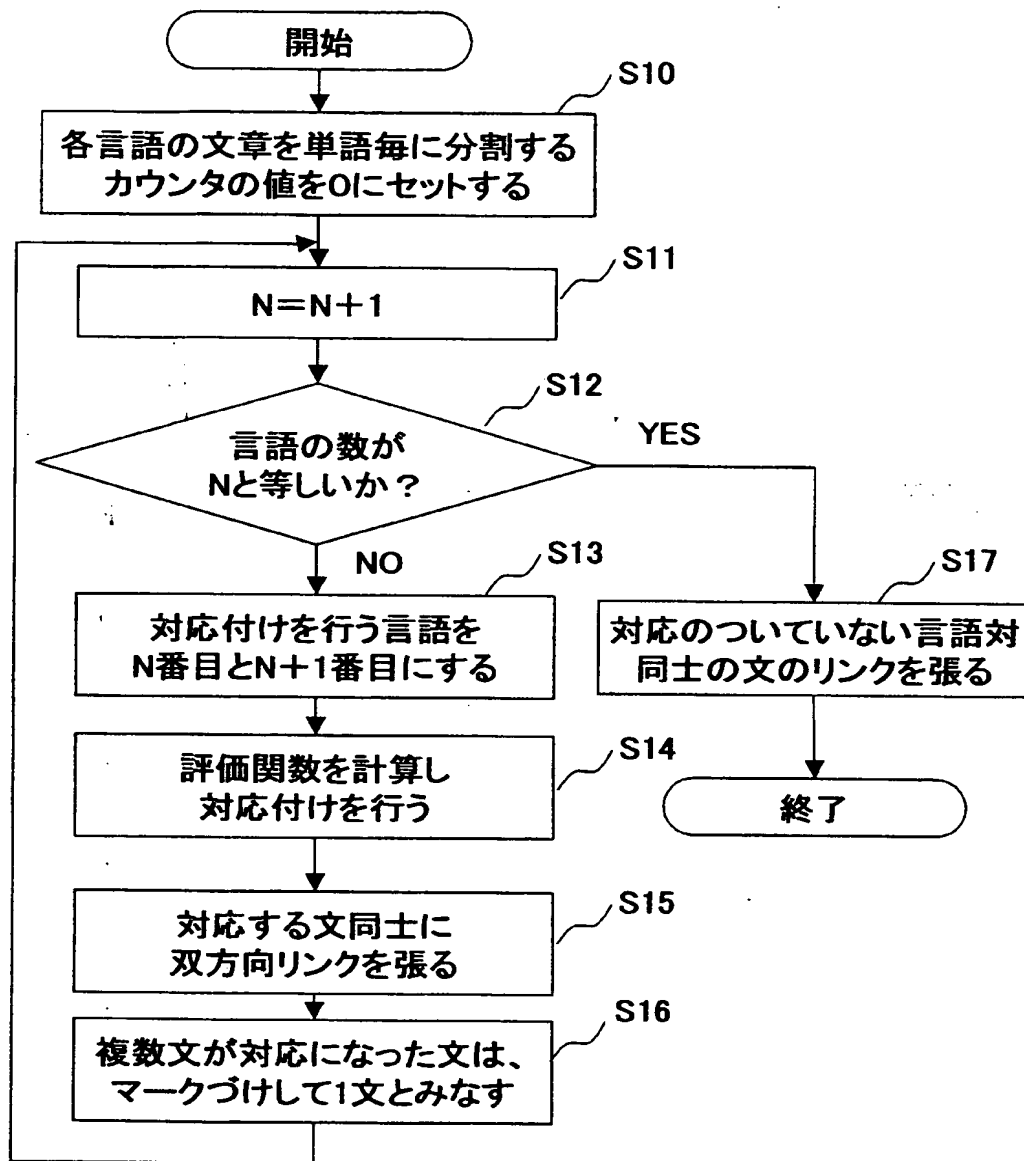
4 2 0 言語類似度データ

【書類名】 図面

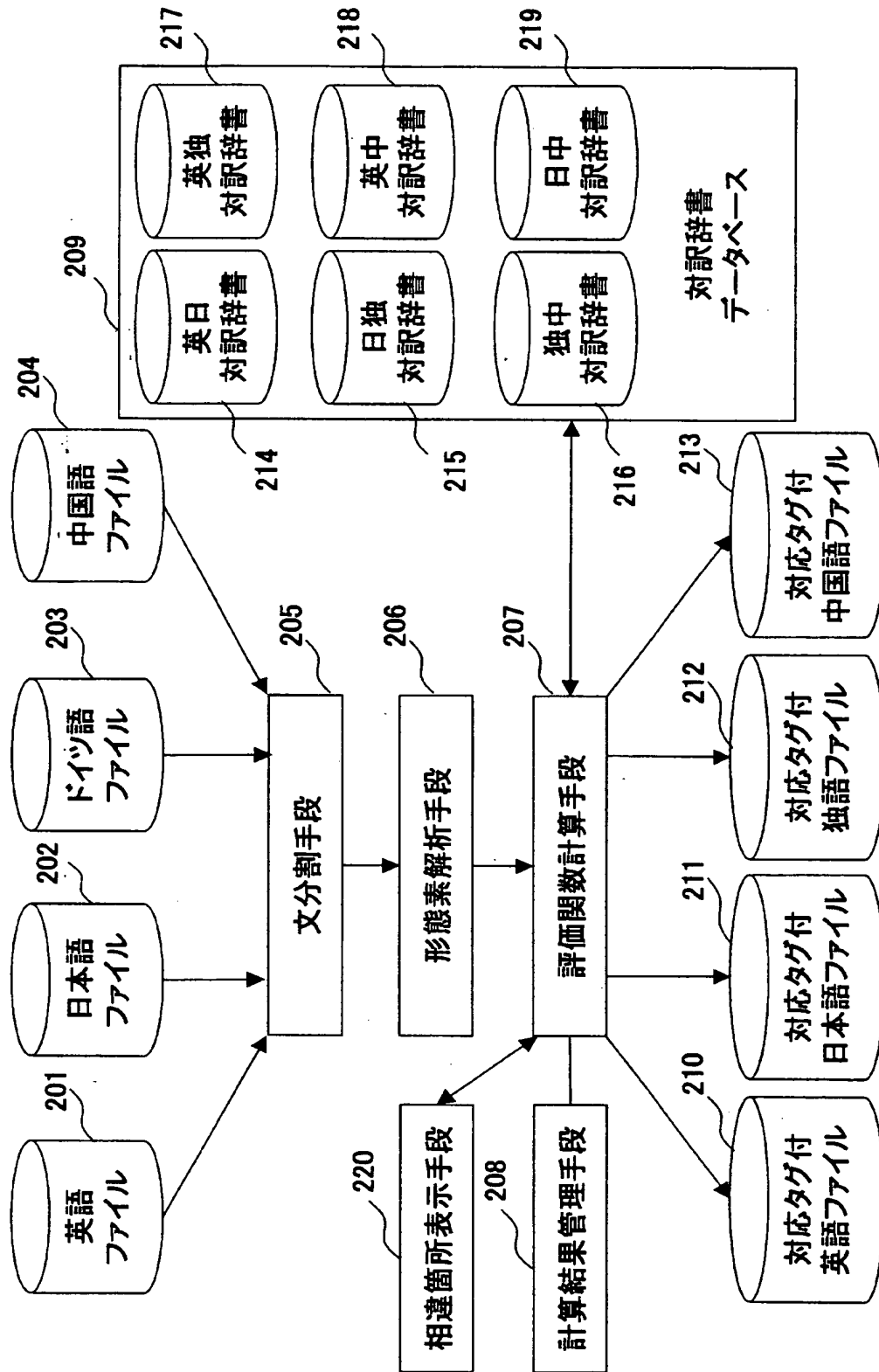
【図 1】



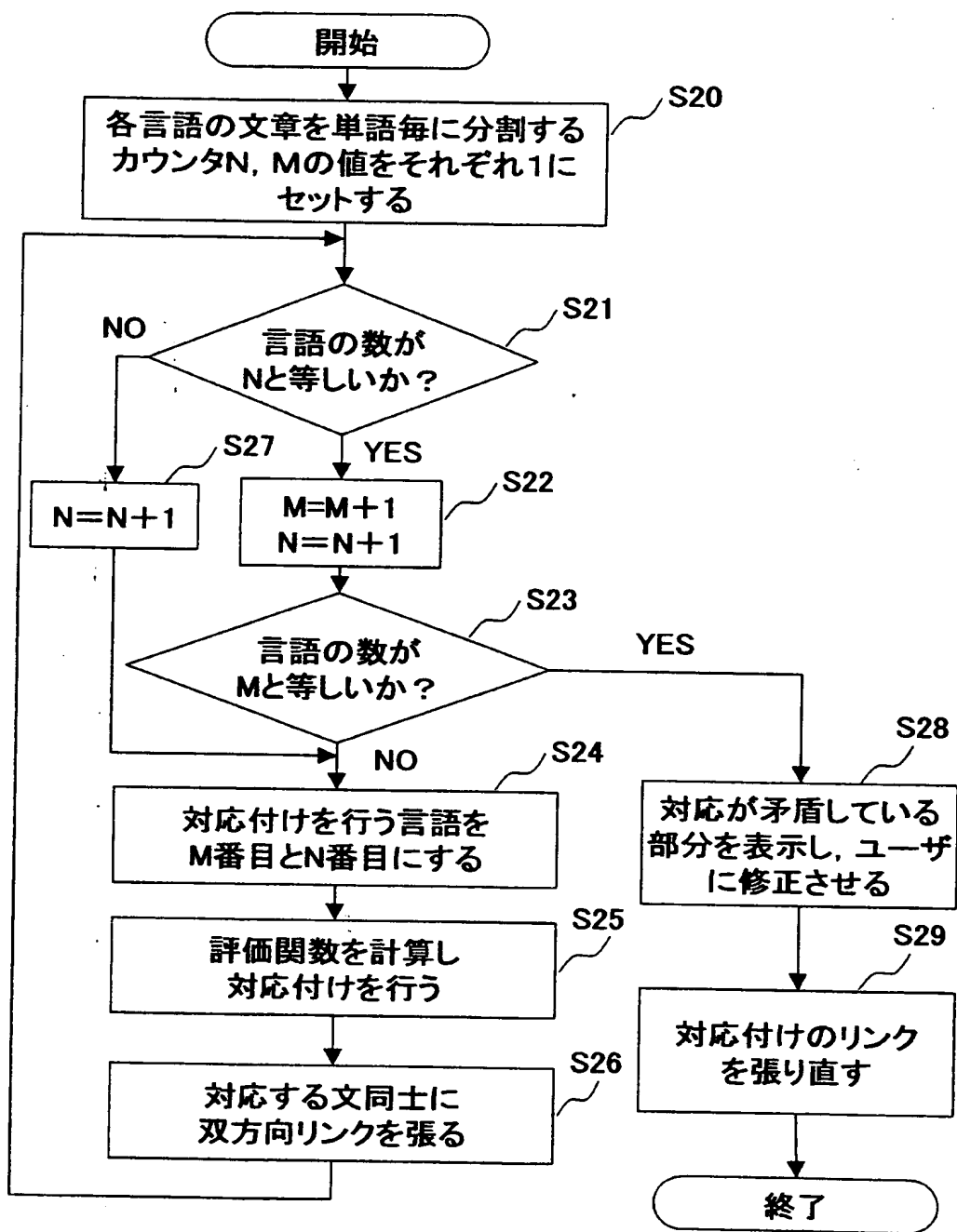
【図2】



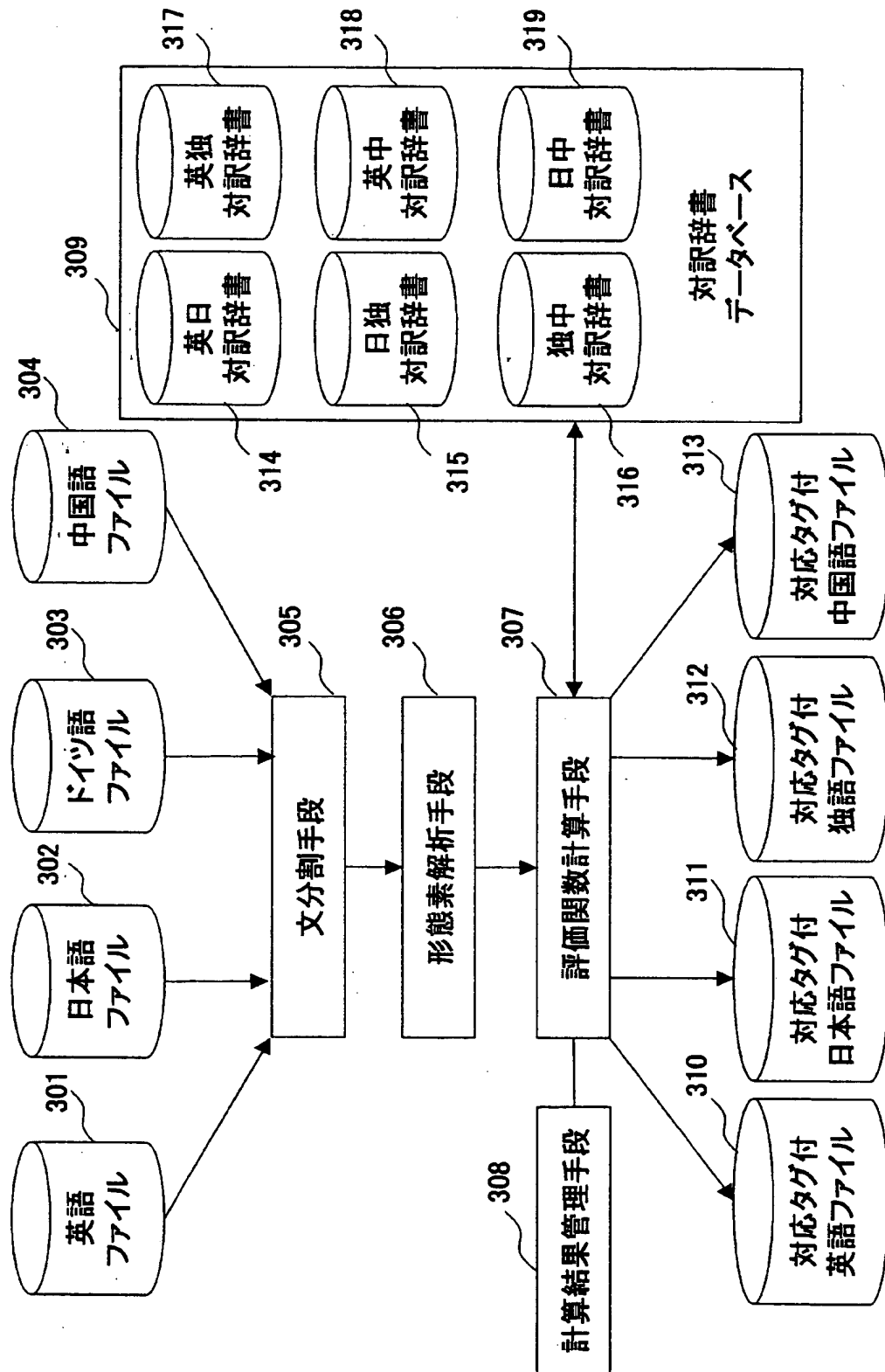
【図 3】



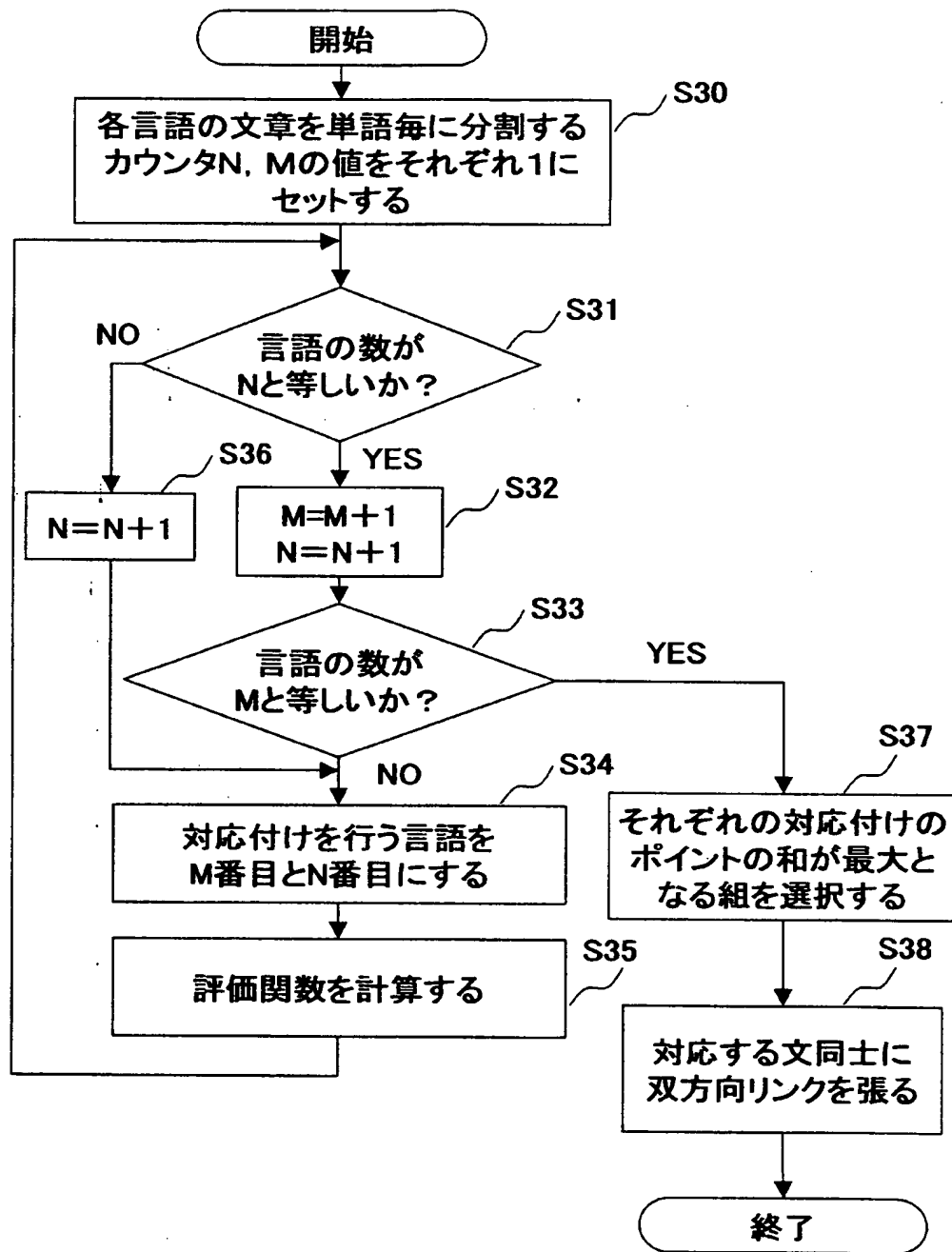
【図 4】



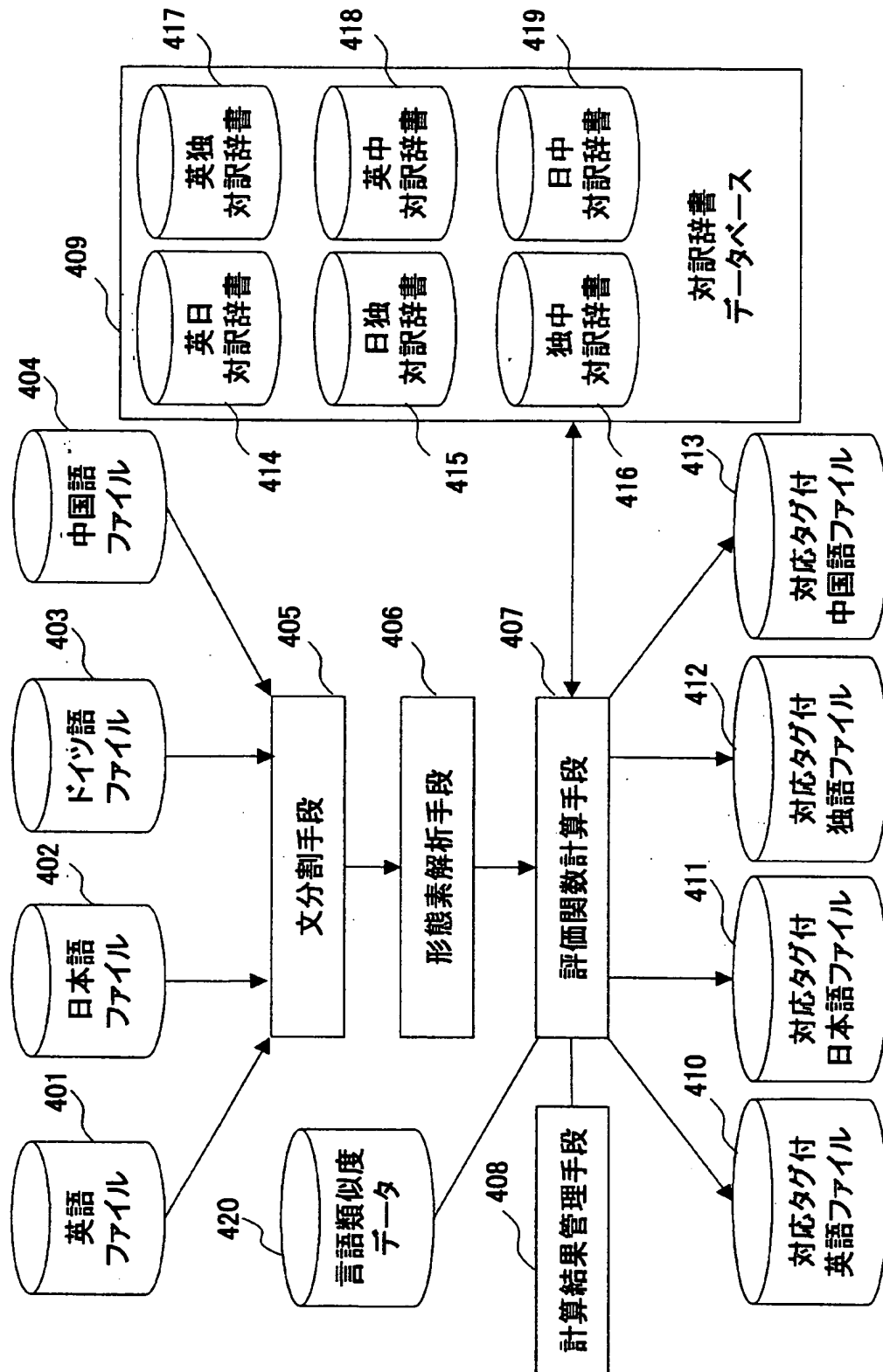
【図 5】



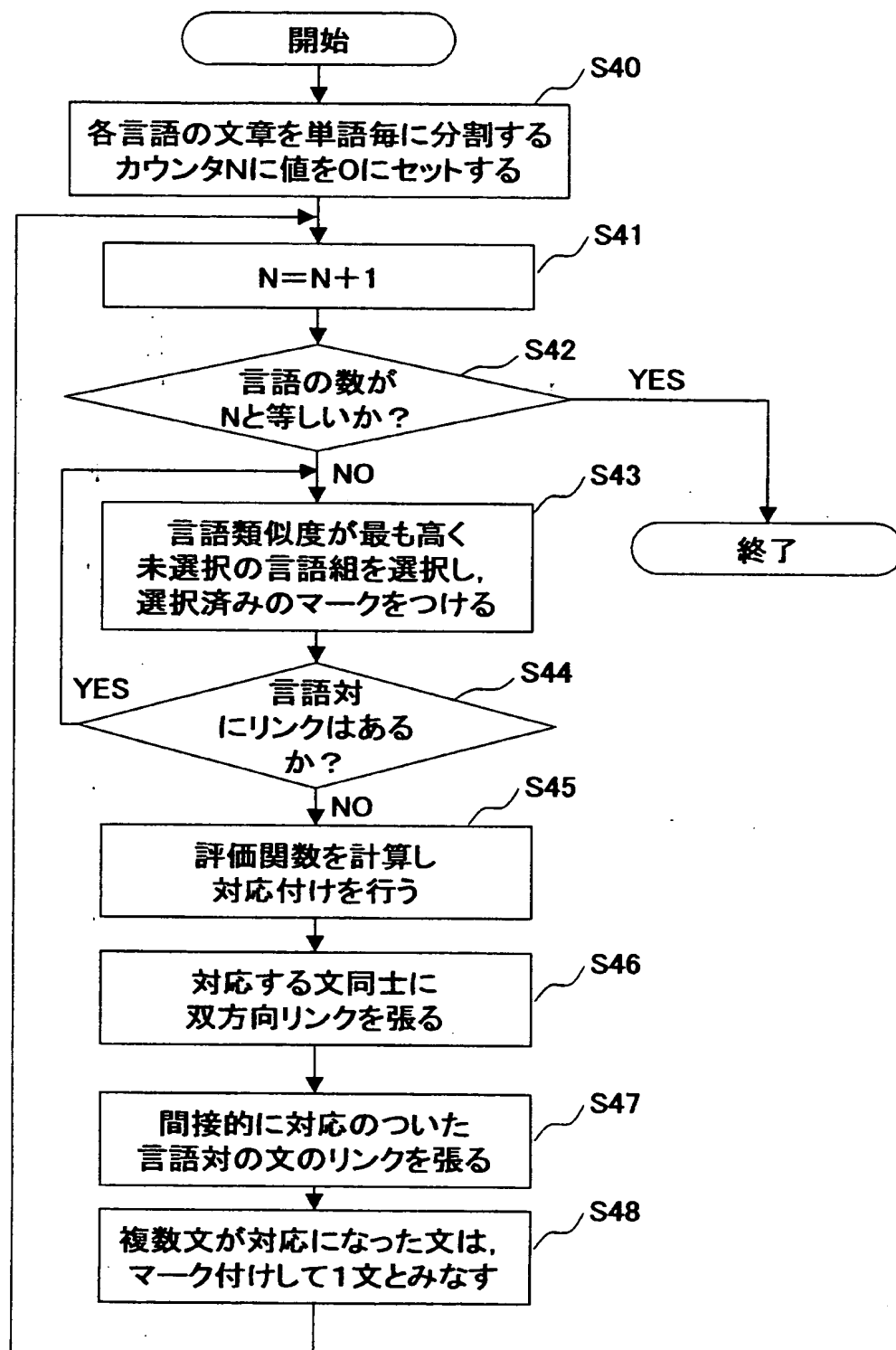
【図 6】



【図 7】



【図 8】



【書類名】 要約書

【要約】

【課題】 複数の言語で構成される文書間の文の対応付けを効率良く行う複数言語文書の対応付けシステムを提供する。

【解決手段】 n 種 (n は 2 以上) の言語の文書を単語毎に分割する形態素解析手段と, n 種の言語の文書のうちの 2 種を選択する手段と, 選択された 2 種の言語文書の評価関数を計算する手段と, 評価結果に応じて n 種の言語の文書を対応付ける手段と, を含むことを特徴とする複数言語文書の対応付けシステム。各言語の文書を単語毎に分割する形態素解析手段は, 各言語の文書を文毎に分割する手段と, 分割された各文をさらに単語毎に分割する手段とからなってもよい。

【選択図】 図 1

特願 2 0 0 2 - 3 4 5 9 8 8

出 願 人 履 歴 情 報

識別番号

[0 0 0 0 0 0 2 9 5]

1. 変 更 年 月 日

1 9 9 0 年 8 月 2 2 日

[変 更 理 由]

新 規 登 録

住 所

東 京 都 港 区 虎 ノ 門 1 丁 目 7 番 1 2 号

氏 名

沖 電 気 工 業 株 式 会 社